The Hagen Matrices Test (HMT)

Timo Heydasch

University of Hagen

Abstract

Intelligence is one of the most central constructs in psychology and is of profound importance for individuals' academic or job achievements and health. Even though a wide range of reliable, valid, and approved intelligence tests exists, there are not many free ones. The Hagen Matrices Test (HMT) introduced in this paper is a free web-based intelligence test focused on reasoning. This study (N = 1,339) presents evidence for the reliability of the HMT. Furthermore, associations with other intelligence tests, self-rated multiple intelligences, self-efficiency related measures, as well as dimensions and facets of personality traits are used to demonstrate the convergent and discriminant validity of the HMT. Associations between the HMT and measures of academic performance were used to demonstrate criterion validity. The HMT can be requested at http://HMT.de.lv.

Keywords: Intelligence, Intelligence Measures, Test Reliability, Test Validity, Hagen Matrices Test

Zusammenfassung

Intelligenz ist eines der bedeutendsten Konstrukte in der Psychologie und auch auf individueller Ebene relevant für den akademischen oder beruflichen Erfolg. Auch wenn eine Vielzahl von reliablen, validen und etablierten Intelligenztests existiert, gibt es nicht viele, die frei verfügbar sind. Der vorgestellte Hagener Matrizen-Test (HMT) ist ein kostenfreier webbasierter Intelligenztest, der die Fähigkeit zum schlussfolgernden Denken misst. Die durchgeführte Studie (N = 1 339) belegt die Reliabilität des HMT. Weiterhin konnte konvergente und diskriminante Validität durch Assoziationen mit anderen Intelligenztests, mit selbsteingeschätzten multiplen Intelligenzen, Maßen der Selbstwirksamkeit bzw. mit Dimensionen und Facetten von Persönlichkeitsmerkmalen belegt werden. Korrelationen des HMT mit Indikatoren akademischen Erfolgs demonstrieren die Kriteriumsvalidität. Kostenfreie Nutzungen des HMT können angefragt werden unter http//:HMT.de.lv.

Schlagworte: Intelligenz, Intelligenztests, Hagener Matrizen-Test

The Hagen Matrices Test (HMT)

It is a fact that empirical research quantifying a psychological construct such as intelligence depends on the measurement of this construct. However, and this is the starting point of this paper, there are not many reliable, valid, and noncommercial intelligence tests. Most intelligence tests are commercial and have to be purchased; this applies not only to the manuals but also to the materials needed to administer the test. To begin to alleviate this shortage, the free-of-cost 20-item web-based Hagen Matrices Test (HMT) was developed, which is theoretically classified primarily according to the Cattell-Horn-Carroll (CHC) model of intelligence (Schneider & McGrew, 2012).

The shortage of free-of-charge intelligence tests is of particular relevance because intelligence is one of the core constructs of psychological research, and it is associated with multiple, diverse, and important life outcomes: “Intelligence predicts important things in life” (Deary, 2012, 648; for a brief review see Deary, 2012). In particular, the impact of intelligence in job-related fields has been demonstrated: Certain jobs tend to be limited to more intelligent people (Harrell, 1946; Harrell & Harrel, 1945) and intelligence is positively associated with training success (Hülsheger, Maier, Stumpp, & Muck, 2006; Salgado, Anderson, Moscoso, Bertua, de Fruyt, & Rolland, 2003; Ziegler, Dietl, Danay, Vogel, & Bühner, 2011) and job performance (Bell, 2007; Hunter & Hunter, 1984; Salgado et al., 2003). In addition, intelligence is connected to academic success (e.g., Poropat, 2009), biological factors such as symmetry (Banks, Batchelor, & McDaniel), brain size (McDaniel, 2005), and sperm quality (Arden, Gottfredson, Miller, & Pierce, 2009). Other results suggest associations between intelligence and mental health, or conversely, with mental diseases such as attention-deficit/hyperactivity disorder (Bridgett & Walker, 2006), schizophrenia (Dickson, Laurens, Cullen, & Hodgins, 2011; Fioravanti, Carlone, Vitale, Cinti, & Clare, 2005), or anorexia nervosa (positively, Lopez, Stahl, & Tchanturia, 2010), and even mortality is associated with intelligence (Roberts, Kuncel, Shiner, Caspi, & Goldberg, 2007). Besides the

large interest in the associations between intelligence and other psychological constructs or real life outcomes researchers have also investigated the nature of intelligence, for example, its heritability (Devlin, Daniels, & Roeder, 1997), developmental aspects (Erdfelder, 1987; Salthouse, 1982), trainability (Klauer & Phye, 2008; te Nijenhuis, van Vianen, & van der Flier, 2007), and first and foremost the number and structure of mental abilities (see below).

The importance of intelligence is impressive, despite the fact that there is no or only little agreement about its definition (Holling, Preckel, & Vock, 2004; Willis, Dumont, & Kaufman, 2011; see also Wasserman, 2012). Intelligence appears to represent a person's mental ability to find or create solutions to problems, whereby a debate was carried out with regard to whether intelligence is one global ability (*g*; sensu Jensen, 1998; Spearman, 1904a) or a composition of different distinguishable mental abilities (sensu Guilford, 1967; Thurstone, 1938). In addition, there was debate about the number of abilities and about how to construct an appropriate hierarchical model to represent narrower and broader abilities and in some cases *g* (e.g., Cattell, 1987; Horn & Noll, 1997; Vernon, 1964). Carroll (1993) brought these debates forward with his meta-analytic study in which he collected and analyzed correlations between intelligence tests to determine the number, contents, and hierarchical structure of human cognitive abilities. Based on his results, Carroll propagated the Three-Stratum (TS) theory that differentiates three hierarchical levels of abilities: "*narrow* (stratum I), *broad* (stratum II), and *general* (stratum III)" (p. 633) abilities. This organizational system and other major aspects of the TS theory were integrated into the CHC model of intelligence (Schneider & McGrew, 2012; see also McGrew, 1997, 2005) which is also based on the Horn-Cattell Gf-Gc theory (Horn & Noll, 1997). The CHC model was introduced by Schneider and McGrew (2012) as taxonomy: On the one hand, this model specifies different abilities, and on the other hand, it organizes these abilities and attempts to explain theoretically "how and why people differ in their various cognitive abilities" (p. 99).

According to Carroll's (1993) analyses, figural matrices tests primarily measure induction: The test taker's task is "to inspect a set of materials and from this inspection induce a rule governing the materials, or a particular or common characteristic of one or more stimulus materials, such as relation or a trend" (Carroll, 1993, p. 211). Schneider and McGrew's (2012) definition of induction is quite similar: "The ability to observe a phenomenon and discover the underlying principles or rules that determine its behavior" (p. 112).

Aside from the abilities of *general sequential reasoning* and *quantitative reasoning* (see Carroll, 1993; Schneider & McGrew, 2012), *induction* is the core aspect of the broader ability of *fluid reasoning* (Gf; Schneider & McGrew, 2012). Schneider and McGrew (2012) define Gf as "…the deliberate but flexible control of attention to solve novel, 'on-the-spot' problems that cannot be performed by relying exclusively on previously learned habits, schemas, and scripts" (p. 111). In addition to this close relation between induction and fluid reasoning, fluid reasoning is closely associated with g if not identical to it (Schneider & McGrew, 2012).

In contrast to the CHC model, other taxonomies, models, and theories of intelligence distinguish the content (i.e., verbal, numeric, or figural) of test materials and abilities in a more prominent way. This differentiation can be found, for example, in the Radex model (Guttman, 1965; Guttman & Levy, 1991), the Structure of Intellect model (Guilford, 1967), the Berlin Model of Intelligence Structure (BIS; Jäger, 1982), and the Hierarchical Protomodel of Intelligence Structure Research (HPI; Liepmann, Beauducel, Brocke, & Amthauer, 2007). As the HMT uses figural matrices, it is obviously a figural test.

In the context of these theoretical assumptions, the construction and validation of the HMT is presented in this paper. To do so, data was collected from test takers and the duration of the test was determined. Furthermore, the HMT items were analyzed as well as the properties of the HMT scores (deviation, internal consistency, retest reliability, associations

with sex and age). Furthermore, the factor structure was explored and the convergent validity (associations with other measures of intelligence) was examined, discriminant validity (correlations with less related or nonrelated psychological constructs such as personality traits), and criterion-related validity (associations with academic success) of the HMT.

Method

Overview

Altogether, four studies were conducted to develop and validate the HMT. The first three were pilot studies designed to assess preliminary versions of the HMT.[1] Based on the results of these pilot studies, the final 20-item version of the HMT was constructed. This final HMT version was administered and validated in the fourth study, the results of which are presented in this paper.

Participants

Students enrolled in a distance B.Sc. Psychology course were recruited via email and the university's online-studies web-page. Students received course credit for their participation. A total of 1,902 students worked on the HMT. After several steps of data cleaning (see below), the sample consisted of 1,339 participants (76% women). Their mean age was 32.2 years (*SD* = 8.97).

Measures

Cognitive abilities.

Hagen Matrices Test (HMT). The HMT consists of three parts: the instructions, 20 matrices,[2] and the presentations of the individual scores. The instructions advise participants to complete 3x3 matrices with one missing field. Test takers have 2 min to choose one of eight presented alternatives of which just one completes a matrix correctly. The fit of the missing pieces results from the matrices' structure which is composed of defined and announced rules: horizontal and/or vertical addition, subtraction and/or varying the positions (rotation or movement) of separate elements. These principles are illustrated by two sample

matrices (see http://ww3.unipark.de/uc/HMT_preview for the instructions and sample matrices). Following the instructions, the 20 items are presented. A time counter informs test takers about the amount of time that has passed for each item. If test takers do not mark an answer within 2 min, the next item is presented. In order to provide the individual test scores in the third part of the HMT, an automatic analysis is performed online during the test. Correct answers are coded 1, and false or missing answers are coded 0. The sum and percentage of correct answers are computed. These statistics and additionally the corresponding IQ score (M = 100, SD = 15) including the 90% confidence interval for IQ are presented individually to the test takers. The IQ scores are estimated on the basis of the comparison between the test scores of the HMT and the IQ scores of the reasoning scale of the Intelligence Structure Test 2000 R (Liepmann et al., 2007) using the equipercentile method according to Angoff (1984). This method allows the comparison of nonequivalent tests with different distributions (Lienert & Raatz, 1998).

Intelligence Structure Test 2000 R (I-S-T 2000 R). The extended German I-S-T 2000 R (Liepmann at al., 2007; see also Beauducel, Liepmann, Horn, & Brocke, 2010) was used to measure different domains of cognitive abilities: reasoning, knowledge, and memory. The scales reasoning and knowledge are calculated by an aggregation of verbal (V), numeric (N), and figural (F) subscales. Each subscale consists of three subtests so that the reasoning and knowledge scores are each based on nine subtests. Additional fluid (g_f) and crystallized (g_c) intelligence are individually calculated as orthogonal factors based on (the statistically dependent) reasoning and knowledge measures. Short-term memory is measured by two subscales with verbal and figural content, respectively. The verbal, numeric, and figural subscales as well as reasoning and memory represent the operationalization of five primary mental abilities as proposed by Thurstone (1938). The factors g_f and g_c correspond to Cattell's (1987) model.

10-Minute Test (10MT). General mental ability was measured with the online version of the 10MT (Hilbig & Musch, 2010; see also Grothe, 2011), which is the adaptation of the paper-and-pencil version (Musch et al., 2009). The 10MT primarily measures *g*. Concurrent validity was demonstrated by its association with other measures of intelligence (see Ostapczuk, Musch, & Lieberei, 2011). The content and structure of the 10MT are similar to the Wonderlic cognitive ability tests (especially the Wonderlic Classic Cognitive Ability Test, formerly the Wonderlic Personnel Test; WPT).

Inventory of self-rated intelligence (ISI). The ISI (Rammstedt & Rammsayer, 2002) is based on the Self-Estimates of Intelligence Questionnaire (e.g., Furnham, 2001) which was constructed to measure multiple dimensions of intelligence according to Gardner (e.g., Gardner, 1993). Rammstedt and Rammsayer added dimensions of cognitive abilities by taking Thurstone's (1938) Primary Mental Abilities into account so that the 11 items of the ISI measure verbal comprehension, word fluency, mathematical intelligence, spatial intelligence, memory, perceptual speed, reasoning, musical intelligence, bodily-kinesthetic intelligence, interpersonal intelligence, and intrapersonal intelligence. In the applied online version, participants had to adjust a button on a scale ranging from *extreme low intelligence* to *extreme high intelligence* to rate their multiple intelligences.

Personality traits.

Positive and Negative Affect Schedule (PANAS). The PANAS (Krohne, Egloff, Kohlmann, & Tausch, 1996; see also Watson, Clark, & Tellegen, 1988) differentiates between positive affectivity and negative affectivity. Trait affectivity was measured by instructing participants to rate their emotions and feelings "*…in general*" according to 20 adjectives.

Big Five Inventory (BFI). The Big Five personality traits extraversion, agreeableness, conscientiousness, neuroticism, and openness were assessed with the 44-item German version of the BFI (Lang, Lüdtke, & Asendorpf, 2001; based on John & Srivastava, 1999).

HEXACO Personality Inventory-Revised (HEXACO-PI-R 100). The HEXACO-PI-R 100 (Lee & Ashton, 2004, 2006; see also www.HEXACO.org) measures the six domain-level traits Honesty-Humility (H), Emotionality (E), eXtraversion (X), Agreeableness (A), Conscientiousness (C), and Openness to Experience (O) according to the HEXACO model of personality (Ashton & Lee, 2007). Each domain-level scale consists of four narrower facet-level scales. In addition to the 96 items belonging to H, E, X, A, C, or O, four items build the facet-level scale altruism. Compared to the traits from the Five Factor model, X and C are similar, O is broadly similar, whereas E and A can be interpreted as rotated factors of the dimensions neuroticism and agreeableness, respectively, and H is a dimension not explicitly included in the Five Factor Model (Ashton, Lee, Marcus, & de Vries, 2007).

Personality-Adjective Scales (PASK5). The PASK5 (Brandstätter, 2010, 2012) are based on the 16 Personality Factor model sensu Cattell (e.g., Cattell, 1957) and was developed according to 16 Personality Factor questionnaires (Brandstätter, 1988; Cattell, Cattell, & Cattell, 1993; Schneewind, Schröder, & Cattell, 1983) measuring warmth, reasoning, emotional stability, dominance, liveliness, rule-consciousness, social boldness, sensitivity, vigilance, abstractedness, privateness, apprehension, openness to change, self-reliance, perfectionism, and tension. The 32 items (two per scale) are presented as 9-point semantic differentials between two contrasting adjectives.

Narcissistic Personality Inventory (NPI). The NPI (Schütz, Marcus, & Sellin, 2004; see also Raskin & Hall, 1979) measures narcissism.[3] The 40-item version was administered in which each item presents two statements, one of which indicates narcissism, with a forced-choice format.

Self-related concepts.

General perceived self-efficacy (GSE). The GSE scale (Schwarzer & Jerusalem, 1995) measures a person's general nonspecific perceived self-efficacy which is a central construct in Bandura's social cognitive theory (Bandura, 1997).

Study-specific self-efficacy (SSSE). The SSSE scale (Schiefele, Moschner, & Husstegge, 2002) is a measure of perceived self-efficacy in the field of studying. The revised seven-item version of the original scale by Jerusalem and Schwarzer (1986) was used.

Self-concept scale (SCS). SC was measured by scales representing academic (nine items), mathematical (six items), and linguistic self-concept (eight items). The items stem from the SMILE project (Schiefele et al., 2002) and are comparable to the corresponding scales from the Self-Description Questionnaire III (Marsh & O'Neill, 1994).

Self-esteem scale (SES). The SE scale (von Collani & Herzberg, 2003) is the German version of Rosenberg's self-esteem scale (Rosenberg, 1965) and a revision of the former adaptation by Ferring and Filipp (1996).

Helplessness. The general helplessness scale (GHELP; short version with seven items) and the study-specific helplessness scale (SSHELP, six items; Jerusalem & Schwarzer, 1986, 2012) were administered as 5-point Likert scales (in contrast to the original 4-point scale). They measure two different aspects of helplessness according to the theory of learned helplessness (Seligman, 1975): perceived general helplessness and perceived helplessness in the field of studying.

Social desirability.

Social Desirability Scale (SDS-17). The SDS-17 (Stöber, 1999) was constructed as the successor to the Marlowe-Crowne Social Desirability Scale (Crowne & Marlowe, 1960; Lück & Timaeus, 1969). It consists of 17 personal behavior statements which the participant has to rate as *true* or *false*.

Balanced Inventory of Desirable Responding (BIDR). The BIDR (Musch, Brockhaus, & Bröder, 2002; according to Paulhus, 1991) identifies desirable responding based on tendencies toward self-deceptive enhancement and impression management

Achievement motivation.

Mehrabian Achievement Risk Preference Scale (MARPS). The MARPS (Mikula, Uray, & Schwinger, 1976, 2012) is an adaptation of the original Achievement Risk Preference Scale (Mehrabian, 1968, 1969) and measures achievement motivation. In contrast to the original instrument, the German version has 20 items (including seven filler items) that are presented to both sexes. The German MARPS is a forced-choice test for which participants have to choose one of two statements.

Achievement Motives Scale (AMS-R). The AMS-R (Lang & Fries, 2006) is the revised short 10-item version of the former version (Dahme, Jungnickel, & Rathje, 1993; Göttert & Kuhl, 1980; as cited in Dahme et al., 1993) which itself is a translation of the original AMS (Gjesme & Nygard, 1970; as cited in Dahme et al., 1993). The AMS-R measures the two dimensions hope of success (HS) and fear of failure (FF).

Achievement Motive Test (AMT). The AMT (Modick, 1977) is a revised and adapted questionnaire. The original is the Dutch Prestatie Motivatie Test (Hermans, 1968; see also Hermans, 2004). It distinguishes between the three scales need for achievement with regard to the future time perspective, debilitating anxiety, and facilitating anxiety.

Criteria. Participants were asked to report different aspects of their academic achievements. They reported their school leaving qualification (SLQ; *Allgemeine Hochschulreife* = 5, *Fachhochschulreife* = 4, *Mittlere Reife* = 3, *Hauptschulabschluss* = 2; no degree = 1), their grade point average (GPA), as well as their last school grades in Mathematics, German, English, Biology, and Arts. Grades in the B.Sc. Psychology courses were also assessed and a B.Sc. Psychology GPA was calculated as the mean of z-standardized grades.

Procedure

The studies were conducted online with the EFS Survey of questback GmbH (see Buchwald, Spoden, Fleischer, & Leutner, 2013). In order to guarantee that just the intended

sample of students would participate, students' access was limited by a password that had been previously revealed only to the students in the distance B.Sc. Psychology course. After entering the correct password, participants were welcomed and given information about the general contents, aim, and expected duration of the study as well as information about voluntary participation and data privacy. They were instructed to answer each question and to work on each task. If a nonresponse occurred, participants were asked to complete their answers (but they were not forced to do so). In contrast to the online assessment, the I-S-T 2000 R was administered as a paper-and-pencil test.

Because of the number and length of administered tests and questionnaires, the data collection was divided into several separate parts. To detect and match the data from individual participants across the different sessions, and to guarantee anonymity at the same time, a 6-digit pseudonymous code was requested.

We performed several steps of data cleaning to ensure protocol validity (Johnson, 2005) concerning the online HMT data. Initially registered hits (N = 3,405) included break-offs (n = 1,384), which primarily resulted from the immediate closing of the browser window after clicking on the public link. Therefore, the first step was to select those cases in which people worked on the HMT and therefore could be identified as test participants (N = 2,021). Those valid trials contained some instances of multiple participations by single participants (n = 119); these were identified by the pseudonymous code and excluded so that each case represented an individual participant (N = 1,902). Finally, participants (n = 563) who took part in at least one of the three HMT preliminary version studies were excluded so that the final sample contained only first-time test takers (N = 1,339) who had no prior HMT test experience.

In the same manner that the HMT data was cleaned, the data for the other measures were prepared: repeated participations on any measure were rejected so that just first-time participants were considered in further analyses.

Results

General Results

The duration of the complete HMT was M = 24.4 min (SD = 12.60, Mdn = 24.1, Range 2.1 to 186.3) so that most participants worked on the HMT for less than half an hour. One part was related to the instructions (M = 4.5 min, SD = 9.05, Mdn = 3.3, Range 0.3 to 163.5) and one to the actual test, which lasted about 20 min (M = 19.9, SD = 8.11, Mdn = 20.4, Range 0.7 to 39.7).

Item difficulties and item-total correlations are presented in Table 1. The HMT contained two easy items with a difficulty of $p > .70$. By contrast, it contained 12 difficult items ($p < .30$). The mean difficulty was M = .37 (SD = .26, Range .10 to .88). The correlation between item position and item difficulty was r = -.94 ($p < .001$). The item-total correlations ranged from r_{it} = .19 to r_{it} = .50.

The mean number of correct responses was M = 7.43 (SD = 3.38); men (M = 8.37, SD = 4.26, N = 347) solved approximately one more matrix correctly than women (M = 7.11, SD = 3.38) with a difference of M_{Δ} = 1.26 (t = 4.98, df = 507.45, $p < .001$, N = 987). In addition to gender effects, age effects were detected as well. The association was r = -.116 ($p < .001$, N = 1,333) indicating that younger participants solved more items correctly. On average, a 21 year younger person answered one more item correctly than an older participant. The regression of the HMT on age was significant for the linear (b = -.048, c = 8.95, R^2 = .013, df_1 = 1, df_2 = 1331, $p < .001$) as well as for the quadratic model (b_1 = .110, b_2 = -.002, k = 6.39, R^2 = .017, df_1 = 2, df_2 = 1330, $p < .001$). This was an effect of d = 0.26 (equivalent to 4 IQ points). Figure 1 shows the graph of the linear and the quadratic regressions of the HMT on age.

Table 1
Item difficulty and item-total correlations

Item	p	r_{it}
1	.88	.30
2	.84	.36
3	.66	.38
4	.67	.33
5	.65	.45
6	.55	.28
7	.56	.38
8	.37	.19
9	.25	.34
10	.29	.42
11	.24	.29
12	.29	.31
13	.21	.38
14	.16	.21
15	.15	.48
16	.13	.39
17	.17	.27
18	.16	.50
19	.12	.30
20	.10	.45

Note. N = 1,339. The standard deviation of dichotomous item is $[p(1\text{-}p)]^{½}$.

To determine reliability, the internal consistency and the retest reliability were computed. The internal consistency was r_{KR8} = .78 according to the Kuder-Richardson Formula 8 (KR8; Kuder & Richardson, 1937). The retest correlation, computed with data from a subsample of 216 participants who worked on the HMT a second time, was r_{tt} = .75 (p < .001). The mean test-retest interval was M = 78 days (SD = 123) with a range of 5 to 388 days. The stability, defined as the retest correlation corrected for attenuation (Spearman, 1910) based on the internal consistency was ρ = .95.

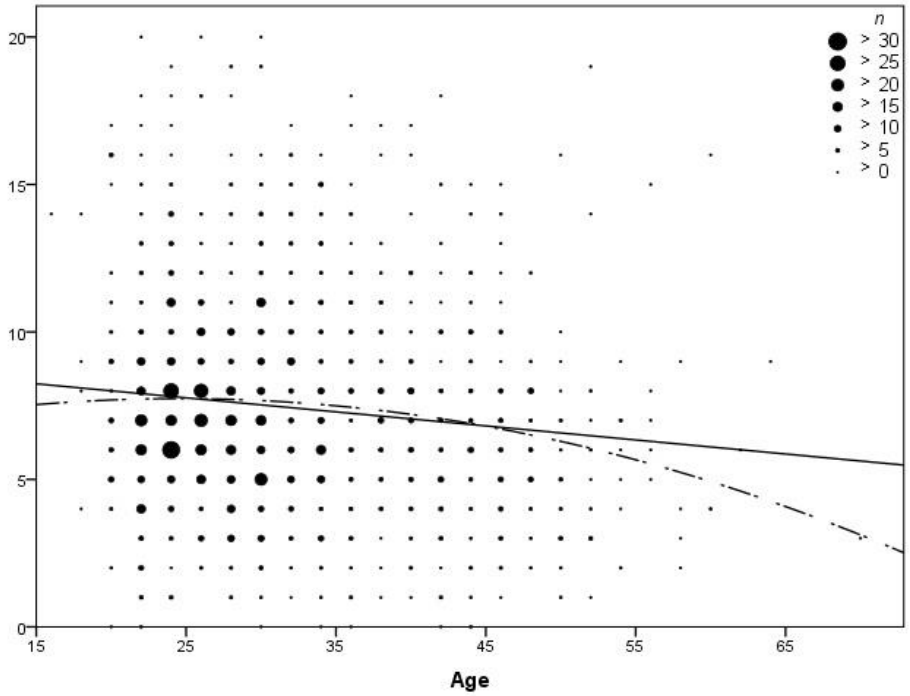

Figure 1. Linear (solid line) and quadratic (intermitted line) regressions of the HMT on age (N = 1,333, larger points indicate a larger subsample).

Factor Analyses[4]

The Kaiser-Meyer-Olkin (KMO; Kaiser & Rice, 1974) value was .697, which was substantially greater than .50 and therefore sufficiently acceptable (Kaiser, 1970; Kaiser & Rice, 1974). Bartlett's test of sphericity led to a rejection of the null hypothesis that the matrix was an identity matrix, $\chi^2(190) = 13.986$, $p < .001$.

To determine the number of factors to retain, a parallel analysis (PA; Horn, 1965; based on O'Connor, 2000) (with 9,000 data sets using principal component eigenvalues), the minimum average partial (MAP) test (Velicer, Eaton, & Fava, 2000; based on O'Connor, 2000), the comparison data (CD) technique (Ruscio & Roche, 2012; using R 2.15.1), and the scree test (Cattell, 1966) were performed. The results of the PA and the MAP test suggested two factors (see Table 2). The two-component solution was supported by the CD technique and by the scree test as well: One obvious break point was located between the second and third components.[5]

Table 2

Parallel analysis (PA) and minimum average partial (MAP) test

	Eigenvalues			MAP test
		PA		
Component	Raw data	*M*	95%	$M(r_{part}^4)$
1	7.010	1.220	1.257	.0029
2	2.212	1.182	1.210	.0004
3	1.118	1.153	1.176	.0003
4	1.082	1.128	1.149	.0004
5	1.010	1.105	1.124	.0010
6	0.889	1.083	1.102	.0017
7	0.813	1.063	1.080	.0038
…	…	…	…	…

Note. 95% = 95th percentile; $M(r_{part}^4)$ = average partial correlation power 4.

Based on the results of the PA, CD technique, scree test, and MAP test, a principal component analysis (PCA) with two predefined factors was conducted (see Table 3). These factors explained 46.1% of the variance (35.1% the first factor). All loadings on Factor 1 were greater than .30 and could be assumed to be substantial for the factor. The factor loadings of Factor 2 varied between a = .53 (Item 1) and a = -.59 (Item 19). Remarkably, the loadings on the second factor decreased continuously. The correlation of the loadings on this factor with item difficulty was r = .85 (p < .001). Thus, Factor 2 appeared to primarily represent the difficulty of the matrices and had to be interpreted as a “spurious” difficulty factor (see McDonald & Ahlawat, 1974), whereas the first factor represented reasoning, the fundamental ability needed to solve matrices. The results of the additional calculation of the measures of sampling adequacy (MSA) confirmed the item characteristics. The range of MSAs ranged from MSA = .36 (Item 8) to MSA = .95 (Item 11).

Table 3
Component matrix of the principal component analysis (PCA) and communalities

	Factors			
Items	1 *Reasoning*	2 *Difficulty*	Communalities	MSA
1	.55	.53	.58	.51
2	.62	.45	.58	.79
3	.58	.33	.45	.63
4	.49	.39	.39	.62
5	.67	.35	.57	.76
6	.42	.31	.27	.89
7	.57	.32	.43	.84
8	.32	-.17	.13	.36
9	.57	-.05	.33	.89
10	.64	.22	.46	.90
11	.51	-.30	.35	.95
12	.50	.05	.25	.62
13	.62	-.03	.38	.68
14	.37	.19	.17	.64
15	.79	-.27	.70	.88
16	.70	-.46	.70	.74
17	.49	-.01	.24	.92
18	.81	-.30	.75	.70
19	.55	-.59	.66	.51
20	.79	-.47	.84	.57

Note. MSA = measures of sampling adequacy.

Validity

The validation of the HMT included correlations with other measures of intelligence, personality, other constructs such as self-efficacy and social desirability, motives, and academic success criteria.

The highest correlations with other measures of intelligence (see Table 4) were found for the reasoning ability measures from the I-S-T 2000 R. The HMT was correlated with r = .57 with general reasoning and with r = .53 with g_f. The correlations with figural and numeric

reasoning ability were r = .51 and r = .50, respectively. Verbal reasoning was not as closely related to HMT (r = .34). The HMT was also correlated with other facets of intelligence at levels ranging from r = .24 for verbal to r = .39 for figural knowledge. General knowledge, g_c, and memory were associated with the HMT at r = .38, r = .30, and r = .28, respectively.

Table 4
Correlations between the HMT and intelligence measures

Variable	*N*	*r*	*KR20*
I-S-T 2000 R	91		
Reasoning		.57***	.93
Verbal		.34***	.77
Numeric		.50***	.93
Figural		.51***	.80
g_f		.53***	
Knowledge		.38***	.85
Verbal		.24*	.69
Numeric		.34***	.65
Figural		.39***	.69
g_c		.30**	
Memory		.28**	.82
10MT	65	.45***	.77
ISI	1332		
Vocabulary		-.04	-
Word fluency		-.08**	-
Numeric		.30***	-
Spatial		.23***	-
Memory		-.06*	-
Perception speed		.01	-
Reasoning		.19***	-
Musical		-.04	-
Physical bodily-kinesthetic		-.06*	-
Interpersonal		-.13***	-
Intrapersonal		-.12***	-

Note. KR20 = Internal consistency according to the Kuder-Richardson Formula 20 (Kuder & Richardson, 1937); I-S-T 2000 R = Intelligence Structure Test 2000 R (Liepmann, Beauducel, Brocke, & Amthauer, 2007); 10MT = 10-Minute Test (Hilbig & Musch, 2010); ISI = Inventory of self-estimated intelligence (Rammstedt & Rammsayer, 2002); g_f = fluid intelligence factor; g_c = crystallized intelligence factor.
* p < .05. ** p < .01. *** p < .001.

The correlation of the HMT with the 10MT was $r = .45$.

The analyses of the self-estimated intelligence scores revealed some divergent results. The HMT was positively correlated with the self-estimated numeric ($r = .30$), spatial ($r = .23$), and reasoning ($r = .19$) abilities and negatively correlated with the self-estimated interpersonal ($r = -.13$) and intrapersonal ($r = -.12$) abilities. All other intelligence measures of the ISI were uncorrelated with the HMT.

The correlations with the personality traits are presented in Table 5. Positive affectivity, negative affectivity, all Big Five dimensions measured by the BFI, and narcissism were not correlated with the HMT. In addition, the HEXACO-PI-R 100 dimensions honesty-humility, agreeableness, and conscientiousness, including the belonging facets, were uncorrelated. There were significant correlations with the dimension emotionality ($r = -.09$), the belongingness facet fearfulness ($r = -.12$), and the facets sociability ($r = -.11$), inquisitiveness ($r = .11$), unconventionality ($r = .08$), and altruism ($r = -.08$). Most traits of the PASK5 were also not correlated with the HMT with the exception of warmth ($r = -.10$), reasoning ($r = -.21$), and openness to change ($r = .12$).

Even though the nine correlations presented above were significant, three of them (emotionality, unconventionality, and altruism from the HEXACO-PI-R 100) were very small at $-.10 < r < .10$, and five (fearfulness, sociability, and inquisitiveness from the HEXACO-PI-R 100, as well as warmth and openness to change from the PASK5) were evaluated as small according to Cohen (1988).

The correlations with the self-related variables differed in their absolute values and directions (see Table 6). The correlations ranged from $r = .36$ (mathematics self-concept) to $r = -.14$. (study-specific helplessness). The two different types of helplessness were negatively correlated with the HMT, and self-efficacy and the self-concepts were positively correlated with the HMT, whereas the more closely the variables were related to reasoning and academic abilities, the higher were the absolute coefficients.

Table 5

Correlations between the HMT and personality traits (part 1)

Variable	*N*	*r*	α
PANAS	587		
Positive Affectivity		.01	.87
Negative Affectivity		-.06	.88
BFI	406		
Extraversion		.02	.88
Agreeableness		.05	.79
Conscientiousness		.00	.85
Neuroticism		-.04	.89
Openness		-.08	.83
HEXACO-PI-R 100	694		
Honesty-Humility		.01	.82
Sincerity		.00	.70
Fairness		.01	.76
Greed Avoidance		.02	.79
Modesty		.01	.67
Emotionality		-.09*	.80
Fearfulness		-.12**	.64
Anxiety		-.03	.70
Dependence		-.06	.71
Sentimentality		-.05	.69
eXtraversion		-.05	.85
Social Self-Esteem		.01	.70
Social Boldness		-.03	.68
Sociability		-.11**	.66
Liveliness		-.04	.74
Agreeableness		.00	.83
Forgivingness		.01	.71
Gentleness		.00	.63
Flexibility		-.06	.50
Patience		.03	.73

Note. Table continues on the next page (part 2).

Table 5
Correlations between the HMT and personality traits (part 2)

Variable		*N*	*r*	α
HEXACO-PI-R 100		694		
Conscientiousness			-.02	.79
Organization			-.07	.67
Diligence			-.04	.70
Perfectionism			.06	.66
Prudence			-.02	.57
Openness to Experience			.04	.75
Aesthetic Appreciation			-.06	.63
Inquisitiveness			.11**	.65
Creativity			.00	.55
Unconventionality			.08*	.42
(Altruism)			-.08*	.57
PASK5		505		
A	Warmth		-.10*	.63
B	Reasoning		.21***	.56
C	Emotional stability		.04	.81
E	Dominance		-.02	.52
F	Liveliness		-.05	.59
G	Rule-consciousness		.01	.47
H	Social boldness		.00	.72
I	Sensitivity		-.02	.70
L	Vigilance		.00	.47
M	Abstractedness		.01	.56
N	Privateness		.00	.12
O	Apprehension		-.06	.70
Q1	Openness to change		.12**	.74
Q2	Self-reliance		.03	.53
Q3	Perfectionism		.02	.69
Q4	Tension		-.07	.77
NPI		576	-.02	.83

Note. PANAS = Positive and Negative Affect Schedule (Krohne, Egloff, Kohlmann, & Tausch, 1996); BFI = Big Five Inventory (Lang, Lüdtke, & Asendorpf, 2001); HEXACO-PI-R 100 = 100-item HEXACO Personality Inventory-Revised (Lee & Ashton, 2004, 2006); PASK5 = Personality-Adjective Scales (Brandstätter, 2010, 2012); NPI = Narcissistic Personality Inventory (Schütz, Marcus, & Sellin, 2004).
* $p < .05$. ** $p < .01$. *** $p < .001$.

Table 6

Correlations between the HMT and the self-related concepts, social desirability, and explicit achievement motivation

Variable	*N*	*r*	α
Self-efficacy			
GSE	548	.10*	.89
SSSE	353	.21***	.90
SCS			
Academic	671	.21***	.85
Mathematic	673	.36***	.93
Linguistic	662	.00	.84
SES	455	.06	.91
Helplessness			
GHELP	508	-.10*	.86
SSHELP	353	-.14**	.86
SDS-17	401	.00	.70
BIDR			
Self-deceptive enhancement	587	.04	.64
Impression management	587	-.05	.70
MARPS	487	.08	.66
AMS-R			
Hope of Success	486	.17***	.86
Fear of Failure	486	-.07	.84
AMT			
Need achievement with regard to future time perspective	464	.04	.88
Debilitating anxiety	464	-.06	.92
Facilitating anxiety	464	.08	.89

Note. GSE = General perceived self-efficacy (Schwarzer & Jerusalen, 1995); SSSE = Study-specific self-efficacy (Schiefele, Moschner, & Husstegge, 2002); SCS = Self-concept scales (Schiefele, Moschner, & Husstegge, 2002); SES = Self-esteem scale (v. Collani & Herzberg, 2003); GHELP = Scale of general helplessness (Jerusalem & Schwarzer, 1986, 2010); SSHELP = study specific helplessness scale (Jerusalem & Schwarzer, 1986, 2010); SDS-17 = Social Desirability Scale (Stöber, 1999); BIDR = Balanced Inventory of Desirable Responding (Musch, Brockhaus, & Bröder, 2002); MARPS = Mehrabian Achievement Risk Preference Scale (Mikula, Uray, & Schwinger, 1976, 2009); AMS-R = Achievement Motives Scale (Lang & Fries, 2006); AMT = Achievement Motive Test (Modick, 1977).

* $p < .05$. ** $p < .01$. *** $p < .001$.

The HMT was unrelated to social desirability: Neither the SDS-17 nor the BIDR scales showed significant correlations. The analyses of the associations with explicit achievement motivation revealed a significant and substantial correlation with the AMS-R hope of success scale. All other achievement motivation scales had a zero correlation with the HMT.

Table 7 provides an overview of the results concerning the associations of the HMT with different academic achievements. There was a slight association with the school-leaving qualification (r = .15). Participants with a higher level of education and as a consequence a longer duration in school, solved more HMT items. In addition, both high school and university GPA were positively associated with the HMT (r = .19 and r = .25, respectively). Not all grades were correlated with the HMT: Grades in the school subjects English (as a foreign language), German, and the arts were unrelated, whereas mathematics (r = .27) and biology (r = .12) were positively associated with the HMT as well as the students' statistics' grades (r = .36) in the psychology course.

The time interval between high-school graduation and test participation in the study was usually 10 years or more. The calculated "retrospective" validity depended on and was attenuated by this interval; therefore, the coefficient for the criterion-related validity for the younger subgroup (age < 24 years, M = 21.67, SD = 1.14) was calculated to get results from a sample which is rather comparable to samples of other test validations. In addition, because not all types of school-leaving qualifications were comparable across all fields, the participants who had the *Abitur* were selected, which characterized the largest group. The results of the analyses with this subsample are also presented in Table 7 (in parentheses). All correlations were higher: The correlations between the HMT and high school GPA, mathematics grades, and biology grades were medium to large (r = .34, r = .45, r = .35, respectively). In the subsample, English grades were also associated (r = .21) with the HMT, but the grades in German and the arts were not.

Table 7
Correlations between the HMT and indicators of academic achievement

Variable	*N*	*r*	
High school[a]			
SLQ[b]	637 (118)	.15***	
GPA[c]	645 (118)	.19***	(.34***)
Mathematics[c]	641 (118)	.27***	(.45***)
English[c]	639 (118)	.07	(.21*)
German[c]	641 (118)	.00	(.08)
Biology[c]	626 (114)	.12**	(.35***)
Arts[c]	610 (113)	-.01	(.12)
B.Sc. Psychology			
GPA[c]	255	.25***	
Statistics[c]	140	.36***	

Note. GPA: Grade point average. SLQ: School-leaving qualification. Results in parentheses were computed on a homogenous subsample of participants younger than 24 years (M = 21.67, SD = 1.14) who all had the same school-leaving qualification (*Abitur*). [a] Participants who did not have a German high-school degree were excluded because of the diverse international coding of degrees. [b]Spearman (1904b) correlations. [c]Grades were recoded so that positive correlations would indicate that higher HMT scores occurred with better grades.
* p <.05. ** p < .01. *** p < .001.

Discussion

The aim of this contribution was to introduce the web-based HMT and to report general characteristics of the items and the scale as well as results concerning its reliability and validity.

General characteristics

The HMT is a relatively short test with a mean duration of less than half an hour. Thus, the test promises an economic procedure to ensure that its acceptance should therefore be relatively high. There was a large standard deviation and a wide range for the testing time. In particular, very short and very long testing times led to the reasonable suspicion of inappropriate participation and invalid measures. In this regard, future investigations may aim to improve the validity of individual protocols (see Johnson, 2005; Kurtz & Parrish, 2001)

and thus to improve the reliability and validity of the HMT, even though these criteria were already shown to be quite satisfactory in the current study (see below).

Although the HMT is fairly short, it is quite challenging. Most items are rather difficult and the mean number of correct answers was quite low. This might result in a stress response in individual participants when they notice that they are having problems identifying the correct answers or when they receive their results. There are two ways to alleviate this critical issue: On the one hand, the Results section of the HMT informs participants that the test is difficult and that other participants in general solve only a few tasks. On the other hand, I highly recommend that the test should be administered only to samples of at least average, or even better, above average intelligence.

Another issue with regard to the difficulty of the test is that sex differences have been found. The sex differences on the HMT are equivalent to 4.65 IQ points, a difference that represents a small to medium effect according to Cohen (1988). Considering the wide range of existing intelligence tests and the diversity of tasks included on them, a wide range of sex differences can be found (for an overview, see Halpern, Beninger, & Straight, 2011) because sex differences vary according to the measured facet or factor, the tasks that are utilized, and the characteristics of the sample. Irwing and Lynn's (2005) meta-analysis is comparable to our own study; they used data from 22 Raven's Progressive Matrices (RPM) studies in which all participants were all university students. The authors found that an effect size of $d = 0.31$ provided the best estimate of the sex differences for students measured with the RPM. Therefore, the HMT reflected valid gender differences and was in this regard fair, or was at least as fair as the RPM.

In contrast to the effects of sex differences, age turned out to play only a minor role. Hertzog (2011) found a correlation of $r = -.40$ in his review of the association between age and fluid intelligence. Results deviated greatly from that age effect such that the expected decline was confirmed, but the magnitude was quite lower. One reason may be found in the

relatively young sample. An accelerated and larger decline is typically found in older samples (Anstey, Hofer, & Luszcz, 2003; Ghisletta, Rabbitt, Lunn, & Lindenberger, 2012). But overall, the regressions of the HMT on age in Figure 1 show a very familiar pattern with an overall linear decline, and in addition, the increasing and then decreasing quadratic curve commonly found in longitudinal (Hertzog, 1989) and cross-sectional studies (see also Erdfelder, 1987; Salthouse, 1982).

The internal consistency and the retest correlation demonstrated the appropriateness of the HMT for group-level analyses (see Lienert & Raatz, 1998). However, analyses at the individual level would be problematic, because the standard error of measurement and the standard error of estimate respectively are quite large at the given level of reliability (estimated by the internal consistency). To say it the other way round: observed individual scores are relatively bad estimated of individuals true scores. For group level analyses this deficit can be countered by aggregation of test scores of larger samples.

PA, the CD technique, and the MAP test suggested a two-component structure. This solution was also supported by the scree test. In determining the number of "true" factors, Kaiser's (1960) Eigenvalue greater than one criterion was ignored, which suggested no fewer than five factors. The MAP test, scree test, and especially PA and DC detect latent dimensions with a high accuracy. By contrast, the eigenvalue greater than one criterion has been found to be less reliable and tends to overestimate the number of components (Ruscio & Roche, 2012; Zwick & Velicer, 1986). Thus, the two-factor solution appears quite appropriate.

The contents of the first factor must be explained from two perspectives. First, matrices have been widely established for measuring intelligence and reasoning ability in particular. Second, the validation coefficients show how the HMT's major component has to be interpreted as a factor that represents induction, reasoning, and fluid intelligence (see below). All items had a substantial loading greater than .30 on the first component; a closer look, however, revealed that some items were less adequate. Items 8 and 14 had loadings of

only a < .40 on the first component. In addition, Item 8 had a very low MSA of .36, whereas the MSAs of the other items were above .50. Future studies will show whether these items should be eliminated and whether the HMT needs to be revised.

We are quite confident about the nature of the second component too. The correlations between the item difficulties and their loadings demonstrated their similarity. Also, the range of item difficulty and skewness of the HMT scale indicated the existence of a difficulty factor (McDonald & Ahlawat, 1974). Thus, regarding its content, the HMT can be conceived of as a unidimensional instrument.

A measure of a psychological construct should show high correlations with other established measures of the same construct. Therefore, the HMT should be highly correlated with the I-S-T 2000 R scales of (figural) reasoning. In fact, the highest correlations were those of the HMT with the I-S-T 2000 R scales of reasoning, g_f, and figural reasoning. These correlations were substantial, significant, and according to Cohen (1988), large. The deviations of the empirical correlation coefficients from perfect may be explained, for example, by the imperfect reliabilities (of both the HMT and the I-S-T 2000 R scales), by individual varying motivations between the test sessions, and by the different methods: non-proctored online (HMT) versus proctored paper-and-pencil assessments (I-S-T 2000 R). Nevertheless, the associations demonstrate the validity of the HMT as a test of intelligence, more specifically of reasoning and induction, and especially of figural reasoning. In sum, these results clearly support the assumption that the HMT is a test of *induction* and GF according to the CHC model of intelligence (Schneider & McGrew, 2012).

In addition, however, the correlation of the HMT with the numeric reasoning I-S-T 2000 R scale (r = .50) was similar in magnitude to the correlations with reasoning, g_f, and figural reasoning. Taking the structure of the matrices into account, the explanation of this finding is quite straightforward: the construction of the matrices was based on the rules of addition and subtraction. This quasi-mathematical principle seems to have had an impact on

the validity of the HMT so that the HMT is also a test of numeric reasoning and thus includes two content areas that are usually considered to be separate (e.g., Guttman, 1965). Transferring this insight to the CHC model of intelligence (Schneider & McGrew, 2012), the HMT has to be characterized as a test of quantitative reasoning: "The ability to reason, either with induction or deduction, with numbers, mathematical relations, and operators" (p. 112). This emphasizes the conceptual relatedness of the HMT to the broader ability of Gf.

The other correlations with measures of intelligence systematically decreased such that the more the aspects of the intelligence measures deviated from the I-S-T 2000 R (figural) reasoning measures, the lower the correlations with the HMT were. These aspects consist of content (verbal vs. figural or numeric), domain of intelligence (e.g., knowledge or memory vs. reasoning), and assessment method (self-rating vs. performance test).

The HMT was unrelated to nearly all personality traits. Even if there were some significant correlations, they were small (or less than small), and such marginal associations should be evaluated as substantial only if these results are replicated. In sum, the HMT was neither substantially nor significantly associated with personality, thus demonstrating its discriminant validity. Only the PASK5 reasoning factor has to be viewed as an associated personality trait. The items of the PASK5 reasoning factor are self-ratings of the individual ability to think (i.e., the speed and easiness of thinking). Therefore, this result is comparable to the self-rated reasoning ability of the ISI variable reasoning and supports convergent validity.

Stern (1911) pointed out that a measure of intelligence should measure intelligence, but naturally is not *only* a measure of intelligence. Other influences can also impact the measure; for example, motivational aspects or fatigue (see also Conrad, 1983). The extant results, however, suggest the independence or just weak associations of the HMT regarding explicit achievement motivation and social desirability.

Intelligence is, among other influences, a decisive factor in academic success (Poropat, 2009). Therefore, the associations with the diverse indicators of academic success confirmed the (criterion-related) validity of the HMT. In particular, the substantial associations with subjects that require inductive abilities such as mathematics and statistics underline the criterion-related validity of the HMT.

Limitations and future directions

The present study is limited in several ways. Perhaps most severely: The results are based on a specific student sample that is far from being representative. This is true for the age and the sex distributions of the sample, but may also be true for additional characteristics of psychology students such as personality traits. Future research should use more diverse or perhaps even representative samples. Additionally, other measures of intelligence should be selected in order to strengthen the construct validity of the HMT.

There are also problematic aspects of the HMT itself: These are its difficulty, its length, and its existence as solely a web-based test. The difficulty is quite challenging. I strongly recommend that the HMT be applied only in studies with well-educated participants, or if conceivable, that the expected samples' mean intellectual abilities be above average. A further restriction is the defined length of the HMT of 20 items: A test version with more items would promise a higher reliability (if a longer test time is acceptable); with a shorter form, the duration could be reduced (if a lower reliability is acceptable); with an adaptive version, a flexible test version could be realized.

Conclusion

In conclusion, the HMT is a sufficiently reliable instrument for measuring intelligence, and reasoning in particular, in group analyses. The extant results on the construct and criterion-based validity are encouraging and justify the use of the HMT. Even if some properties of the test limit its possible applications, it could become a useful research tool,

especially because the HMT is free. For further information and test requests, visit http://HMT.de.lv.

References

Angoff, W. H. (1984). *Scales, norms, and equivalent scores.* Princeton, NJ: Educational Testing Service.

Anstey, K. J., Hofer, S. M., & Luszcz, M. A. (2003). A latent growth curve analysis of late-life sensory and cognitive function over 8 years: Evidence for Specific and Common Factors Underlying Change. *Psychology and Aging, 18*(4), 714-726. doi:10.1037/0882-7974.18.4.714

Arden, R., Gottfredson, L. S., Miller, G., & Pierce, A. (2009). Intelligence and semen quality are positively correlated. *Intelligence, 37*(3), 277-282. doi:10.1016/j.intell.2008.11.001

Ashton, M. C., & Lee, K. (2007). Empirical, theoretical, and practical advantages of the HEXACO model of personality structure. *Personality and Social Psychology Review, 11*(2), 150-166. doi:10.1177/1088868306294907

Ashton, M. C., Lee, K., Marcus, B., & De Vries, R. E. (2007). German lexical personality factors: Relations with the HEXACO Model. *European Journal of Personality, 21*(1), 23-43. doi:10.1002/per.597

Bandura, A. (1997). *Self-Efficacy: The Exercise of Control*. New York, NY: Freeman.

Banks, G. C., Batchelor, J. H., & McDaniel, M. A. (2010). Smarter people are (a bit) more symmetrical: A meta-analysis of the relationship between intelligence and fluctuating asymmetry. *Intelligence, 38*(4), 393-401. doi:10.1016/j.intell.2010.04.003

Beauducel, A., Liepmann, D., Horn, S., & Brocke, B. (2010). *Intelligence Strukture Test (IST).* Göttingen: Hogrefe.

Bell, S. T. (2007). Deep-level composition variables as predictors of team performance: A meta-analysis. *Journal of Applied Psychology, 92*(3), 595–615. doi:10.1037/0021-9010.92.3.595

Brandstätter, H. (1988). Sechzehn Persönlichkeits-Adjektivskalen (16PA) als Forschungsinstrument anstelle des 16PF. *Zeitschrift für experimentelle und angewandte Psychologie, 35*, 370-391.

Brandstätter, H. (2010). *Persönlichkeits-Adjektiv Skalen (PASK5). Manual.* University of Linz.

Brandstätter, H. (2012). Persönlichkeits-Adjektiv Skalen (PASK5). Fünf Faktoren Modell. In A. Glöckner-Rist (Ed.), *Zusammenstellung sozialwissenschaftlicher Items und Skalen*. ZIS version 15.00. Bonn: GESIS.

Bridgett, D. J., & Walker, M. E. (2006). Intellectual functioning in adults with ADHD: A meta-analytic examination of full scale IQ differences between adults with and without ADHD. *Psychological Assessment, 18*(1), 1–14. doi:10.1037/1040-3590.18.1.1

Buchwald, F., Spoden, C., Fleischer, J., & Leutner, D. (2013). Verzweigte Lernumgebungen und Tests mit EFS Survey 8. *Diagnostica, 59*(2), 113-117. doi:10.1026/0012-1924/a000080

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge, England: University Press.

Cattell, R. B. (1957). *Personality and motivation structure and measurement.* New York, NY: World Book.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1*(2), 245-276. doi:10.1207/s15327906mbr0102_10

Cattell, R. B. (1987). *Intelligence: Its structure, growth, and action*. New York: Elsevier Science

Cattell, R. B., Cattell, A. K., & Cattell, H. E. (1993). Sixteen Personality Factor Questionnaire, Fifth Edition. Champaign, IL: Institute for Personality and Ability Testing.

Cattell, R.B. (1987). *Intelligence: Its structure, growth, and action*. Amsterdam: Elsevier.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd edition). Hillsdale, NJ: Lawrence Erlbaum Associates.

Conrad, W. (1983). Intelligenzdiagnostik. In K. Groffmann & L. Michel (Eds.), *Intelligenz- und Leistungsdiagnostik* (pp. 104-201). Göttingen: Hogrefe.

Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology, 24*(4), 349-354. doi:10.1037/h0047358

Dahme, G., Jungnickel, D., & Rathje, H. (1993). Güteeigenschaften der Achievement Motives Scale (AMS) von Gjesme und Nygard (1970) in der deutschen Übersetzung von Göttert und Kuhl—Vergleich der Kennwerte norwegischer und deutscher Stichproben. *Diagnostica, 39*(3), 257-270.

Deary, I. J. (2012). Intelligence. *Annual Review of Psychology, 6*(3), 453-482. doi:10.1146/annurev-psych-120710-100353

Devlin, B., Daniels, M., & Roeder, K. (1997). The heritability of IQ. *Nature, 388*(6641), 468–471.

Dickson, H., Laurens, K. R., Cullen, A. E., & Hodgins, S. (2012). Meta-analyses of cognitive and motor function in youth aged 16 years and younger who subsequently develop schizophrenia. *Psychological Medicine, 42*(4), 743–755. doi:10.1017/S0033291711001693

Emmons, R. A. (1987). Narcissism: Theory and measurement. *Journal of Personality And Social Psychology, 52*(1), 11-17. doi:10.1037/0022-3514.52.1.11

Erdfelder, E. (1987). *Die Entwicklung psychometrischer Intelligenz über die Lebensspanne. Aspekte eines allgemeinpsychologischen Zugangs*. Frankfurt am Main: Lang.

Ferring, D., & Filipp, S.-H. (1996). Messung des Selbstwertgefühls: Befunde zu Reliabilität, Validität und Stabilität der Rosenberg-Skala. *Diagnostica, 43*(3), 284-292.

Fioravanti, M., Carlone, O., Vitale, B., Cinti, M. E., & Clare, L. (2005). A meta-analysis of cognitive deficits in adults with a diagnosis of schizophrenia. *Neuropsychology Review*, *15*(2), 73–95. doi:10.1007/s11065-005-6254-9

Furnham, A. (2001). Self-estimates of intelligence: Culture and gender difference in self and other estimates of both general (g) and multiple intelligences. *Personality and Individual Differences*, *31*, 1381-1405.

Gardner, H. (1993). *Multiple intelligences: The theory in practice*. New York, NY: Basic Books.

Ghisletta, P., Rabbitt, P., Lunn, M., & Lindenberger, U. (2012). Two thirds of the age-based changes in fluid and crystallized intelligence, perceptual speed, and memory in adulthood are shared. *Intelligence*, *40*(3), 260-268. doi:10.1016/j.intell.2012.02.008

Grothe, D. (2011). *Kann man Intelligenz in zehn Minuten messen? - Testtheoretische Analyse und Normierung eines zehnminütigen Online-Verfahrens zur Intelligenzdiagnostik*. (Unpublished bachelor‘s thesis) University of Mannheim.

Guilford, J.P. (1967). *The nature of human intelligence*. New York, NY: McGraw-Hill.

Guttman, L. (1965). A faceted definition of intelligence. In R. Eiferman (Ed.), *Studies in psychology, scripta hierosolymitana* (Vol. 14 pp. 166-181). Jerusalem: Hebrew University.

Guttman, L., & Levy, S. (1991). Two structural laws for intelligence tests. *Intelligence*, *15*(1), 79-103. doi:10.1016/0160-2896(91)90023-7

Halpern, D. F., Beninger, A. S., & Straight, C. A. (2011). Sex differences in intelligence. In R. J. Sternberg, S. Kaufman (Eds.), *The Cambridge handbook of intelligence* (pp. 253-272). New York, NY US: Cambridge University Press. doi:10.1017/CBO9780511977244.014

Harrell, T. W. (1946). Army General Classification Test results for air forces specialists. *Educational and Psychological Measurement, 6,* 341-349 DOI: 10.1177/001316444600600304

Harrell, T. W., & Harrell, M. S. (1945). Army General Classification Test scores for civilian occupations. *Educational and Psychological Measurement, 5,* 229-239. doi: 10.1177/001316444500500303

Hermans, H. J. M. (1968). *Prestatie Motivatie Test (PMT).* Amsterdam: Swets & Zeitlinger

Hermans, H. J. M. (2004). *Prestatie Motivatie Test (PMT).* Amsterdam: Pearson Assessment and Information B.V.

Hertzog, C. (1989). Influences of cognitive slowing on age differences in intelligence. *Developmental Psychology, 25*(4), 636-651. doi:10.1037/0012-1649.25.4.636

Hertzog, C. (2011). Intelligence in adulthood. In R. J. Sternberg, S. Kaufman (Eds.), *The Cambridge handbook of intelligence* (pp. 174-190). New York, NY US: Cambridge University Press. doi:10.1017/CBO9780511977244.010

Hilbig, B. E., & Musch, J. (2010). *Online version of the 10 Minutes Test*. Unpublished test, University of Mannheim.

Holling, H., Preckel, F., & Vock, M. (2004). *Intelligenzdiagnostik*. Göttingen: Hogrefe.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30*(2), 179-185. doi:10.1007/BF02289447

Horn, J. L., & Noll, J. (1997). Human cognitive capabilities: Gf-Gc theory. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests and issues* (pp. 53-91). New York: Guilford.

Hülsheger, U. R., Maier, G. W., Stumpp, T., & Muck, P. M. (2006). Vergleich kriteriumsbezogener Validitäten verschiedener Intelligenztests zur Vorhersage von Ausbildungserfolg in Deutschland. *Zeitschrift für Personalpsychologie, 5*(4), 145–162. doi:10.1026/1617-6391.5.4.145

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96*(1), 72-98. doi:10.1037/0033-2909.96.1.72

Irwing, P., & Lynn, R. (2005). Sex differences in means and variability on the progressive matrices in university students: A meta-analysis. *British Journal of Psychology, 96*(4), 505–524. doi:10.1348/000712605X53542

Jäger, A. O. (1982). Mehrmodale Klassifikation von Intelligenzlestungen: Experimentell kontrollierte Weiterenwicklung eines deskriptiven Intelligenzstrukturmodells. *Diagnostica, 28*(3), 195-225.

Jensen, A. R. (1998). *The g factor. The science of mental ability*. Westport, CT: Praeger Pubes.

Jerusalem, M., & Schwarzer, R. (1986). Fragebogen zur Erfassung von Hilflosigkeit. In R. Schwarzer (Ed.), *Skalen zur Befindlichkeit und Persönlichkeit* (Forschungsbericht 5). Berlin: FU.

Jerusalem, M., & Schwarzer, R. (2012). Dimensionen der Hilflosigkeit. In A. Glöckner-Rist (Ed.), *Zusammenstellung sozialwissenschaftlicher Items und Skalen. ZIS Version 15.00*. Bonn: GESIS.

John, O. P., & Srivastava, S. (1999). The Big Five Trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research (2nd ed.)* (pp. 102-138). New York, NY: Guilford Press.

Johnson, J. A. (2005). Ascertaining the validity of individual protocols from Web-based personality inventories. *Journal of Research in Personality, 39*(1), 103–129. doi:10.1016/j.jrp.2004.09.009

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. Educational and Psychological Measurement, 20, 141-151.

Kaiser, H. F. (1970). A second generation little jiffy. *Psychometrika, 35*(4), 401-415. doi:10.1007/BF02291817

Kaiser, H. F., & Rice, J. (1974). Little Jiffy, Mark IV. *Educational And Psychological Measurement, 34*(1), 111-117. doi:10.1177/001316447403400115

Klauer, K. J., & Phye, G. D. (2008). Inductive reasoning: A training approach. *Review of Educational Research, 78*(1), 85–123. doi:10.3102/0034654307313402

Krohne, H.-W., Egloff, B., Kohlmann, C. W., & Tausch, A. (1996). PANAS - Positive and Negative Affect Schedule - deutsche Fassung. *Diagnostica, 42*(2), 139-156.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2*, 151-2160. doi:10.1007/BF02288391

Kurtz, J. E., & Parrish, C. L. (2001). Semantic response consistency and protocol validity in structured personality assessment: The case of the NEO-PI-R. *Journal of Personality Assessment, 76*(2), 315–332. doi:10.1207/S15327752JPA7602_12

Lang, F. R., Lüdtke, O., & Asendorpf, J. B. (2001). Testgüte und psychometrische Ä quivalenz der deutschen Version des Big Five Inventory (BFI) bei jungen, mittlalten und alten Erwachsenen. *Diagnostica, 47*(3), 111-121. doi:10.1026//0012-1924.47.3.111

Lang, J. B., & Fries, S. (2006). A revised 10-item version of the Achievement Motives Scale: Psychometric properties in German-speaking samples. *European Journal of Psychological Assessment, 22*(3), 216-224. doi:10.1027/1015-5759.22.3.216

Lee, K., & Ashton, M. C. (2004). Psychometric properties of the HEXACO personality inventory. *Multivariate Behavioral Research, 39*(2), 329-358. doi:10.1207/s15327906mbr3902_8

Lee, K., & Ashton, M. C. (2006). Further assessment of the HEXACO Personality Inventory: Two new facet scales and an observer report form. *Psychological Assessment, 18*(2), 182-191. doi:10.1037/1040-3590.18.2.182

Lienert, G. A., & Raatz, U. (1998). *Testaufbau und Testanalyse*. Weinheim: PVU.

Liepmann, D., Beauducel, A., Brocke, B. & Amthauer, R. (2007). *Intelligenz-Struktur-Test 2000 R (I-S-T 2000 R)*. Göttingen: Hogrefe.

Lopez, C., Stahl, D., & Tchanturia, K. (2010). Estimated intelligence quotient in anorexia nervosa: a systematic review and meta-analysis of the literature. *Annals of General Psychiatry, 9*:40. doi:10.1186/1744-859X-9-40

Lück, H. E., & Timaeus, E. (1969). Skalen zur Messung Manifester Angst (MAS) und sozialer Wünschbarkeit (SDS-E und SDS-CM). *Diagnostica, 15*, 134-141.

Marsh, H. W., & O'Neill, R. (1984). Self Description Questionnaire III: The construct validity of multidimensional self-concept ratings by late adolescents. *Journal of Educational Measurement, 21*(2), 153-174. doi:10.1111/j.1745-3984.1984.tb00227.x

McDaniel, M. (2005). Big-brained people are smarter: A meta-analysis of the relationship between in vivo brain volume and intelligence. *Intelligence, 33*(4), 337–346. doi:10.1016/j.intell.2004.11.005

McDonald, R. P., & Ahlawat, K. S. (1974). Difficulty factors in binary data. *British Journal Of Mathematical And Statistical Psychology*, *27*(1), 82-99. doi:10.1111/j.2044-8317.1974.tb00530.x

McGrew, K. S. (1997). Analysis of the major intelligence batteries according to a proposed comprehensive Gf-Gc framework. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 151-179). New York: Guilford.

McGrew, K. S. (2005). The Cattell–Horn–Carroll theory of cognitive abilities. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (2nd ed., pp. 136–181). New York: Guilford Press.

Mehrabian, A. (1968). Male and female scales of the tendency of achieve. Educational and Psychological Measurement, 28(2), 493-502. doi:10.1177/001316446802800235

Mehrabian, A. (1969). Measures of achieving tendency. *Educational and Psychological Measurement, 29*(2), 445-451. doi:10.1177/001316446902900222

Mikula, G., Uray, H., & Schwinger, T. (1976). Die Entwicklung einer deutschen Fassung der Mehrabian Achievement Risk Preference Scale. *Diagnostica, 22,* 87-97.

Mikula, G., Uray, H., & Schwinger, T. (2012). Leistungsmotivation. In A. Glöckner-Rist (Ed.), *Zusammenstellung sozialwissenschaftlicher Items und Skalen. ZIS Version 15.00*. Bonn: GESIS.

Modick, H. E. (1977). Ein dreiskaliger Fragebogen zur Erfassung des Leistungsmotivs : Bericht über eine deutschsprachige Weiterentwicklung des Prestatie Motivatie Test. *Diagnostica, 23*(4), 298-321.

Musch, J., Brockhaus, R., & Bröder, A. (2002). Ein Inventar zur Erfassung von zwei Faktoren sozialer Erwünschtheit. Diagnostica, 48(3), 121-129. doi:10.1026//0012-1924.48.3.121

Musch, J., Ostapczuk, M., Hilbig, B. E., Auer, T.S., Brandt, M., Cüpper, L., Erdfelder, E., & Undorf, M. (2009). *10-Minuten-Test*. Unveröffentlichter Test, Universität Düsseldorf.

O'Connor, B. P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavior Research Methods, Instruments & Computers, 32*(3), 396-402. doi:10.3758/BF03200807

Ostapczuk, M., Musch, J., & Lieberei, W. (2011). Der „Analytische Test“: Validierung eines neuen eignungsdiagnostischen Instruments zur Erfassung von schlussfolgerndem Denken. *Zeitschrift für Arbeits- Und Organisationspsychologie, 55*(1), 1-16. doi:10.1026/0932-4089/a000031

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P.R. Shaver, & L.S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (17-59). New York, NY: Academic Press.

Poropat, A. E. (2009). A meta-analysis of the five-factor model of personality and academic performance. *Psychological Bulletin, 135*(2), 322-338. doi:10.1037/a0014996

Rammstedt, B., & Rammsayer, T. (2002). Die Erfassung der selbsteingeschätzten Intelligenz: Konstruktion, teststatistische Überprüfung und erste Ergebnisse des Inventars zur selbsteingeschätzten Intelligenz (ISI). *Zeitschrift für Differentielle und Diagnostische Psychologie, 23*,435-446.

Raskin, R., & Hall, C. S. (1981). The Narcissistic Personality Inventory: Alternate form reliability and further evidence of construct validity. *Journal of Personality Assessment, 45*(2), 159-162. doi:10.1207/s15327752jpa4502_10

Raskin, R., & Terry, H. (1988). A principal-components analysis of the Narcissistic Personality Inventory and further evidence of its construct validity. *Journal of Personality And Social Psychology, 54*(5), 890-902. doi:10.1037/0022-3514.54.5.890

Roberts, B. W., Kuncel, N., Shiner, R., N., Caspi, A., & Goldberg, L. R. (2007). The power of personality: The comparative validity of personality traits, socio-economic status, and cognitive ability for predicting important life outcomes. *Perspectives in Psychological Science, 2*, 313-345. doi:10.1111/j.1745-6916.2007.00047.x

Rosenberg, M. (1965). *Society and the adolescent selfimage*. Princeton, NJ: University Press.

Ruscio, J., & Roche, B. (2012). Determining the number of factors to retain in an exploratory factor analysis using comparison data of known factorial structure. *Psychological Assessment, 24*(2), 282-292. doi:10.1037/a0025697

Salgado, J. F., Anderson, N., Moscoso, S., Bertua, C., de Fruyt, F., & Rolland, J. (2003). A meta-analytic study of general mental ability validity for different occupations in the European Community. *Journal of Applied Psychology, 88*(6), 1068-1081. doi:10.1037/0021-9010.88.6.1068

Salthouse, T. A. (1982). *Adult cognition. An experimental psychology of human aging*. New York, NY: Springer.

Schiefele, U., Moschner, B., & Husstegge, R. (2002). *Skalenhandbuch SMILE-Projekt* (unveröffentlichtes Manuskript). Bielefeld: Universität.

Schneewind, K. A., Schröder, G., & Cattell, R. B. (1983). *Der 16-Persönlichkeits-FaktorenTest (16PF). Testmanual.* Bern: Huber.

Schneider, W., & McGrew, K. S. (2012). The Cattell-Horn-Carroll model of intelligence. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (3rd ed.) (pp. 99-144). New York, NY US: Guilford Press.

Schütz, A., Marcus, B., & Sellin, I. (2004). Die Messung von Narzissmus als Persönlichkeitskonstrukt: Psychometrische Eigenschaften einer Lang- und einer Kurzform des Deutschen NPI (Narcissistic Personality Inventory). *Diagnostica*, *50*(4), 202-218. doi:10.1026/0012-1924.50.4.202

Schwarzer, R., & Jerusalem, M. (1995). Generalized Self-Efficacy scale. In J. Weinman, S. Wright, & M. Johnston (Eds.), *Measures in health psychology: A user's portfolio. Causal and control beliefs* (pp. 35-37). Windsor, England: NFER-NELSON

Seligman, M. E. P. (1975). *Helplessness: On depression, development and death.* San Francisco: Freeman.

Spearman, C. (1904a). "General intelligence" Objectively determined and measured. *The American Journal of Psychology, 15,* 201–292.

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3*, 271–295.

Spearman, C. C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology, 15*(1), 72-101. doi:10.2307/1412159

Stern, W. (1911). *Differentielle Psychologie*. Barth: Leipzig.

Stöber, J. (1999). Die Soziale-Erwünschtheits-Skala-17 (SES-17): Entwicklung und erste Befunde zu Reliabilität und Validität. *Diagnostica, 45*(4), 173-177. doi:10.1026//0012-1924.45.4.173

te Nijenhuis, J., van Vianen, A. E., & van der Flier, H. (2007). Score gains on g-loaded tests: No g. *Intelligence, 35*(3), 283–300. doi:10.1016/j.intell.2006.07.006

Thurstone, L. L. (1938). *Primary and mental abilities.* Chicago: University Press.

Uebersax, J. S. (2007). TetMat 1.0.3. retrieved from http://john-uebersax.com/stat/tetra.htm

Velicer, W. F., Eaton, C. A., & Fava, J. L. (2000). Construct explication through factor or component analysis: A review and evaluation of alternative procedures for determining the number of factors or components. In R. D. Goffin, E. Helmes (Eds.), *Problems and solutions in human assessment: Honoring Douglas N. Jackson at seventy* (pp. 41-71). New York, NY: Kluwer Academic/Plenum Publishers.

Vernon, P. E. (1964). *The structure of human abilities*. London: Methuen.

von Collani, G., & Herzberg, P. (2003). Eine revidierte Fassung der deutschsprachigen Skala zum Selbstwertgefühl von Rosenberg. *Zeitschrift Für Differentielle Und Diagnostische Psychologie*, *24*(1), 3-7. doi:10.1024//0170-1789.24.1.3

Wasserman, J. D. (2012). A history of intelligence assessment: The unfinished tapestry. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues (3rd ed.)* (pp. 3-55). New York, NY US: Guilford Press.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology, 54*(6), 1063-1070. doi:10.1037/0022-3514.54.6.1063

Willis, J. O., Dumont, R., & Kaufman, A. S. (2011). Factor-analytic models of intelligence. In R. J. Sternberg & S. Kaufman (Eds.), *The Cambridge handbook of intelligence* (pp. 39-57). New York, NY US: Cambridge University Press. doi:10.1017/CBO9780511977244.004

Ziegler, M., Dietl, E., Danay, E., Vogel, M., & Bühner, M. (2011). Predicting training success with general mental ability, specific ability tests, and (un)structured interviews: A

meta-analysis with unique samples. *International Journal of Selection and Assessment*, *19*(2), 170–182. doi:10.1111/j.1468-2389.2011.00544.x

Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin*, *99*(3), 432-442. doi:10.1037/0033-2909.99.3.432

Author Note

Timo Heydasch, Lehrgebite Psychologische Methodenlehre, Diagnostik und Evaluation, FernUniversität in Hagen, Germany.

Correspondence conceerning this article should be adressed to Timo Heydasch, FernUniversität in Hagen, Universitätsstr. 33, 58084 Hagen, Germany

E-mail: timo.heydasch@fernuni.hagen.de

Footnotes

[1] The results of the first three studies will not be presented because only the revised and final version of the fourth study is relevant for future applications of the HMT.

[2] Courtesy of Lutz Hornke.

[3] The narcissism subscales were not taken into account because of a lack of agreement concerning the number of subscales and their contents and also because of insufficient reliability: Cronbach's alpha for the 17 subscales (according to Emmons, 1987, Raskin & Terry, 1988, and Schütz et al., 2004) was found to range from $\alpha = .43$ to $\alpha = .76$ ($M_\alpha = .58$, $SD_\alpha = .11$).

[4] The tetrachoric item intercorrelations (calculated with TetMat 1.0.3; Uebersax, 2007) were used in the factor analyses.

[5] The decisive break point can be reconstructed using the reported raw data Eigenvalues in Table 2.